Linking Pictographs to Synsets: Sclera2Cornetto

Vincent Vandeghinste, Ineke Schuurman
Centre for Computational Linguistics
University of Leuven
vincent@ccl.kuleuven.be, ineke@ccl.kuleuven.be

Abstract

Social inclusion of people with Intellectual and Developmental Disabilities can be promoted by offering them ways to independently use the internet. People with reading or writing disabilities can use pictographs instead of text. We present a resource in which we have linked a set of 5710 pictographs to lexical-semantic concepts in Cornetto, a Wordnet-like database for Dutch. We show that, by using this resource in a text-to-pictograph translation system, we can greatly improve the coverage comparing with a baseline where words are converted into pictographs only if the word equals the filename.

Keywords: Pictographic communication, social inclusion, Wordnet

1. Introduction

The importance of the digital society in various aspects of our lives is undeniable. Allowing people with cognitive disabilities to independently use the internet can increase their quality of life, by reducing social isolation (Dawe, 2006; Davies et al., 2001; Newell et al., 2002). Augmentative and Alternative Communication (AAC) assists people with severe communication disabilities to be more socially active in interpersonal interaction, learning, education, community activities, employment, volunteering, and care management.

Picture-based or pictographic communication systems are a form of AAC technology that is based on the use of graphics, such as drawings, pictographs and symbols. These include systems for AAC like Blissymbolics[1], PCS[2], Beta[3], and Sclera[4]. There are estimates that between 2 and 5 million people in the European Union could benefit from pictographic communication as a means of written communication, and there is an acute need for such communication interfaces enabling social contact for people with cognitive disabilities. These interfaces should provide ease of use, configuration, and flexibility in different situations for users with different (dis)abilities (Keskinen et al., 2012). It is in this light that we put the resources described in this paper at the availability of the general public.

WAI-NOT[5] is a Flemish non-profit organisation enabling ICT for children and young people with a mental disability (IDD - Intellectual or Developmental Disability). They developed a website which is adjusted to different intellectual levels. Pictographs and auditory support are used on this website wherever possible, and it is possible to send emails with the help of pictographs through an adjusted email client. For visual support either Beta or Sclera pictures can be used, depending on the user's profile. The user chooses an addressee by selecting a picture in his address book. A message can contain short texts and/or Beta/Sclera-pictures. This communication environment raises the need for improvements in inter-pictograph-language communication, creating translation tools between Dutch, Sclera, and Beta.

We present related work concerning picture-based communication systems in section 2. In section 3. we present resources that allow improvements in communication between users of different pictograph systems, and communication between pictograph systems and natural language. To allow the proper use of pictographic languages for online communication we need some kind of *machine translation* (MT) between natural language text and these pictographic languages (and vice versa). In section 4. we briefly present our approach to translating natural language text into pictographs and show how the resources presented in this paper improve both precision and recall compared to the baseline system. Section 5. draws conclusions and describes our plans for the future.

2. Related work

Pictographic communication has grown from local initiatives of which some have scaled up to larger communities. Across Europe, many pictographic communication systems are in place, but they use no or only a very limited amount of linguistic knowledge to appropriately disambiguate lexical ambiguities, which can lead to wrong conversions into pictographs (Vandeghinste, 2012) or to the conversion into multiple pictographs per word, one for each sense of the word. An application of the latter can be seen on http://www.widgit.com.

Only few works related to the task of translating texts for pictographic communication can be found in the literature. Mihalcea and Leong (2008) describe a system for the automatic construction of pictorial representations for simple sentences and show that the understanding, which can be achieved using visual descriptions, is similar to those of target language texts obtained by means of machine translation. They automatically collected a picture set, which was validated through crowd sourcing. They used WordNet (Miller, 1995) as a lexical resource, but it seems that they did not use the WordNet relations between concepts, as we do.

Goldberg et al. (2008) learn how to improve understanding

[1] http://www.blissymbolics.org/

[2] http://www.mayer-johnson.com/category/symbols-and-photos

[3] http://www.betavzw.be

[4] http://www.sclera.be

[5] http://www.wai-not.org

Figure 1: Pictographs in everyday life

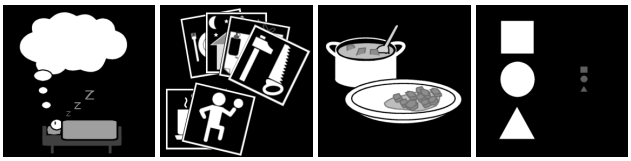

Figure 2: Simplex Sclera pictographs for *"dream", "pictographs", "stew", and "large"*

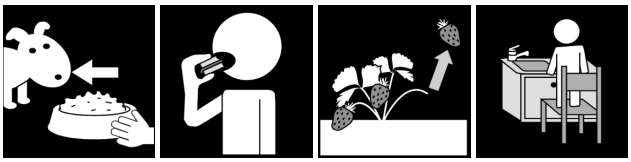

Figure 3: Verb-object pictographs: *"feed the dog", "eat a sandwich", "pick strawberries"* or more complex *"stand on a chair in front of the sink"*

of a sequence of pictographs by conveniently structuring its representation after identifying the different roles which the phrases in the original sentence play with respect to the verb. They use structured semantic role labelling for this.
Joshi et al. (2006) describe an unsupervised approach for automatically adding pictures to a story, extracting semantic keywords from a story and searching an annotated image database. However, they do not try to translate the entire story.
A resource we certainly have to mention is ImageNet (Deng et al., 2009), a large-scale ontology of images linked to the WordNet structure, aiming to populate the majority of the Wordnet synsets. The images in ImageNet seem to be mostly photographs, and are therefore less suitable for communication aids for the cognitively challenged.

3. The resources

After introducing Sclera, the pictograph-set used in this paper, we present Sclera2Cornetto, a resource linking the Sclera pictographs to synonym sets in Cornetto. Cornetto[6] is a lexical-semantic database which is linked to the EuroWordNet[7] grid and to the SUMO ontology,[8] consisting of 118 000 synonym sets (synsets) which are linked to each other through several relationships such as hyponymy, meronymy, and antonymy. The words that are in the Cornetto database are either verbs, nouns, adjectives, or adverbs. It is freely available for non-commercial use from the Dutch HLT centre.[9]
Additionally, we also present Dutch2Sclera, a small dictionary, linking Dutch words that do not appear in Cornetto with pictographs.

3.1. Sclera

Sclera is a large set of mainly black-and-white pictographs. Originally these were used as directives, just like the pictographs we are confronted with in everyday life (as shown in Figure 1).
When people are not able to write and/or read fluently, pictographs may provide a solution. Therefore schools and institutes for people with IDD have since long been using pictographs to guide their pupils and residents.
There are currently over 13000 Sclera pictographs and new pictographs are created every month upon user request. These pictographs are freely available as .png files with a filename indicating their meaning in Dutch, English, French, and Spanish.[10] As shown in Figure 2, they can represent simple concepts corresponding to single Dutch words, but often they represent more complex concepts corresponding, for instance, to a verb and its objects (Figure 3), to two or more nouns (Figure 4), or to nouns and prepositional phrases (Figure 5). There are mainly only pictographs for content words and hardly any pictographs for prepositions or adverbs.
Such pictographs may contain very strict instructions, as shown in Figure 6.
Although Sclera mainly contains black-and-white pictographs, some of them are green (indicating that something is permitted or approved) while others are red (indicating a ban or disapproval). In some others another color is used for contrast or to indicate the color itself. A ban or disapproval may also be expressed by a (red) cross through the pictograph.

As mentioned above, Sclera was originally used as a means to communicate directives to its IDD-users (pupils, residents) with as few pictographs as possible. However, the last decade more and more attention is being paid to the communicative needs of people with IDD, the keyword being *social inclusion*. These users are also entitled to participate in the modern, digital world, by sending e-mails, chatting with friends, and using social networks, among other things. Pictographs are now used in a broader context, i.e., they are on a par with natural languages such as Dutch and English, similar to sign languages.
As a baseline we take the Dutch-to-Sclera system as it was

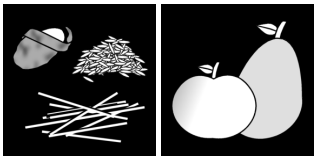

Figure 4: Pictographs with two or more nouns: *"potatoes, rice and pasta"; "pear and apple"*

[6] http://tst-centrale.org/producten/lexica/cornetto/7-56
[7] http://www.illc.uva.nl/EuroWordNet/
[8] http://www.ontologyportal.org/
[9] http://tst-centrale.org/
[10] In this paper we only refer to our work for Dutch.

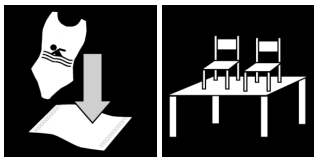

Figure 5: Pictographs with nouns and prepositional phrases: *"swimming suit in towel"; "chairs on table"*

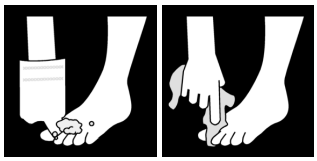

Figure 6: Some pictographs with instructions for personal hygiene: *"wash between toes with soap"* and *"dry between your toes"*

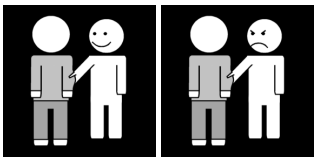

Figure 8: Some pictographs on ways to draw someone's attention: *"draw attention positive"* and *"draw attention negative"*

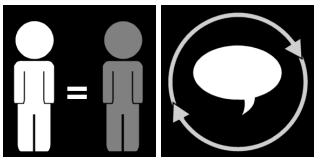

Figure 9: Abstract concepts: *"equal opportunities"* and *"repeat"*

in 2012, before we started working on improving text to pictograph conversion. At that time a sentence like *Ik kom naar huis* was converted as shown in Figure 7.
We are now treating Sclera as a language, albeit a simplified one. Note that this does not imply that it is simple. Pictograph-languages are learned, they do not come naturally. It is for instance hard to *understand* the concepts of the pictographs in Figures 8 and 9 without learning that these pictures stand for these concepts.
When treating messages in Sclera as expressed in natural language (instead of an ad-hoc complex of pictographs), some characteristics of this natural language should be mentioned:

- no articles
- no possessive pronouns
- no inflection
- no tenses
- few auxiliaries (mainly 'to be')
- mostly the same pictograph for singular and plural

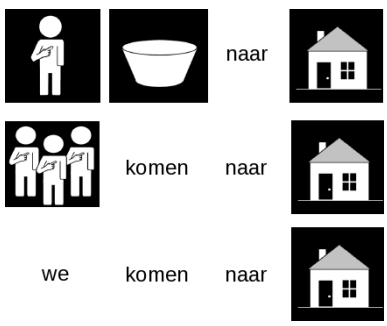

Figure 7: Literal conversion (before 2012): words that do not function as filenames (minus .png) remain untranslated

In some cases this is due to the fact that the concepts involved are hard to put into pictographs (like determiners, inflection of a verb), or because the pictographs mainly express the concept expressed by the lemma. In some cases it is not the lemma that is associated with the pictograph, but, for example, the plural form (when the singlar is lacking).

3.2. Linking Sclera with Cornetto

We have manually linked a subset of 5710 Sclera pictographs to Cornetto synsets. A tool was built which took each Sclera pictograph and checked the Cornetto database to see whether there was an entry with the same name as the filename (without the `.png` extension). If not, the annotator could select one of the senses of the entry. If this was not the case, the annotator could enter a synonym and then select the appropriate sense, or she could tell the annotation tool to connect the pictograph to multiple synsets, providing lemmas that have these synsets as words. For each of these lemmas the appropriate sense was chosen by the annotator.
As these pictographs sometimes depict complex concepts, they can be linked to one or to more synsets indicating that their meaning combines the meanings of the synsets. In these cases we have identified one of the synsets as the *head* synset, indicating that the other linked synsets are in some kind of dependency relation with the head synset. Table 1 presents the distribution of the synsets per linked pictograph. In cases where the pictograph meaning was not reflected by one or more synsets, we often (240 times for simplex pictographs) have linked the pictograph to the synset of its hyperonym.
Sclera2Cornetto consists of a database table with the following columns:

- *lemma:* the name of the pictograph
- for simple pictographs
  - *synset:* synset identifier matching Sclera pictograph

Nr of synsets	Frequency of links between Sclera pictographs and synsets
1	2689
2	2416
3	559
4	42
5	3
6	1

Table 1: Distribution of Sclera pictographs over number of synsets

token	lemma	tag	picto
en	en	VG(neven)	plus.png
	sneeuwen		sneeuw.png
	hallo		hallo-zeggen-2.png
	groet		hallo-zeggen-2.png
	groeten		hallo-zeggen-2.png

Table 2: Some entries in the Dutch2Sclera dictionary

  - *relation:* whether the synset is synonym/hyperonym of pictograph

- for complex pictographs
  - *head:* synset identifier of head
  - *headrel:* relation of synset to pictograph (synonym/hyperonym)
  - *dependent:* comma-separated list of synset identifiers for dependents
  - *deprel:* comma-separated list of relations (synonym/hyperonym) of synsets for each dependent

3.3. The Dutch2Sclera dictionary for Dutch words not covered by Cornetto

We also make our Dutch2Sclera dictionary table available, consisting of 372 entries linking Dutch words straight to Sclera pictographs. This table contains token, lemma, part-of-speech tag, and picto columns, allowing underspecification, cf Table 2. The tagset used is Van Eynde (2005).
Currently, the Dutch-to-Sclera translation system described in section 4. uses this dictionary to distinguish between words where we have pictographs for different meanings of the word, such as *kind* (child) sense: son or daughter) and sense: youngsters, as shown in Figure 10, or *dag* meaning either 'hello' or 'day', cf Figure 11. One of the things on our to-do list is to implement proper word sense disambiguation.

4. Using the resources to translate Dutch into Sclera

We have built a text to pictograph translation system that is used by the WAI-NOT online AAC platform, which allows people who are not able to read and write to communicate through the internet. In this section we briefly report on this system, showing how the presented resources improve the precision and recall in converting Dutch text into Sclera

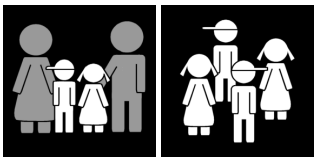

Figure 10: *"Child (son, daughter)"* vs *"child (youngster)"*

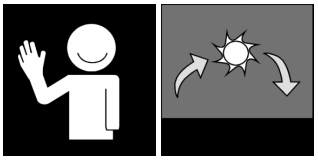

Figure 11: *"Hello"* vs *"day"*, homonyms in Dutch

pictographs. An overview of the architecture of this system is presented in Figure 12.
The input text first undergoes a shallow linguistic analysis (section 4.1.): tokenisation, part-of-speech tagging, sentence splitting, word-based spelling correction for unknown words, separable verb detection and lemmatisation.
In a second stage the synsets of the lemmas of the words are retrieved from Cornetto (section 4.2.).
In a third stage the input sentence is translated into pictographs (section 4.3.).

4.1. Shallow Linguistic Analysis

The first step that we apply is tokenization, splitting of all the punctuation signs from the words, apart from the hyphen/dash and the apostrophe, using a rule-based tokenizer.
The next step concerns word-based spelling correction, as a lot of the messages contain spelling mistakes. These spelling mistakes cannot be considered typing mistakes, but are real errors against the Dutch spelling. Therefore we implemented automatic spelling correction based on the OpenTaal[11] lexicon . For every word that is not in this lexicon, we check all variants with one deletion, one insertion or one substitution. For all the variants present in the OpenTaal lexicon, we take the one that most frequently occurs in the 80 million word corpus. This is currently a 1-gram model. In future versions we might consider higher order versions, if it is deemed necessary.
Then we apply part-of-speech tagging. We use HunPos (Halcsy et al., 2007), a trigram-based open source tagger similar to TnT (Brants, 2000), using the D-Coi tagset (Van Eynde, 2005), trained on the SoNaR corpus (Oostdijk et al., 2013).
As the system is intended to translate e-mail messages for mentally challenged people, messages tend to be short, and mostly consist of only one sentence. Nevertheless, some of the messages contain more than one sentence, so we apply sentence detection, as the translation engine works sentence based.
Dutch contains separable verbs. These are verbs that have a lexical core and a separable particle. In some syntactic situtations the core and the particle are written as one

[11] http://www.opentaal.org

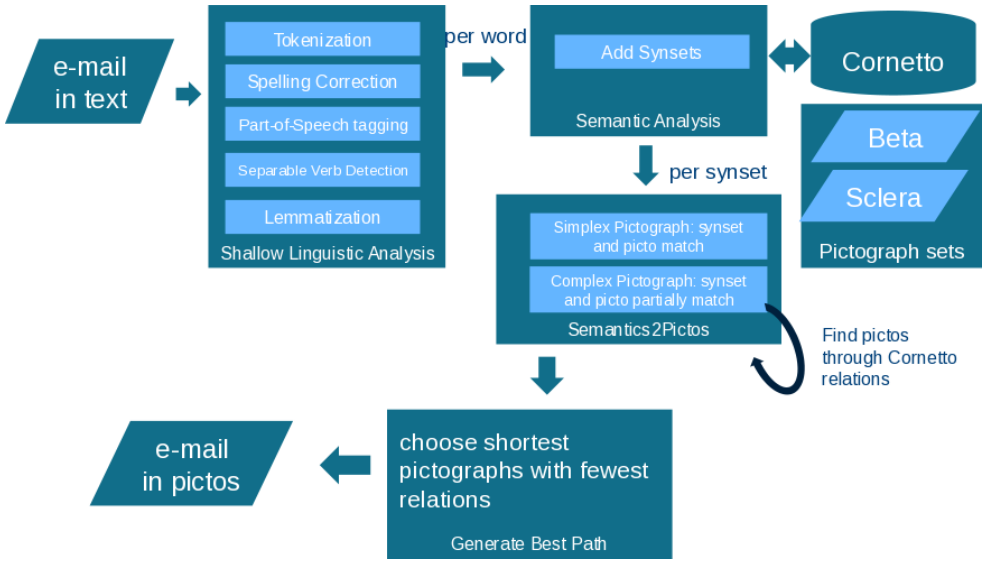

Figure 12: Architecture of the text to pictograph translation system

word, while in other situations they are written separately. Particles can have different part-of-speech tags, according to the tagset we use (Van Eynde, 2005). The most frequent part-of-speech tags for particles are final prepositions `VZ(fin)`, such as in verbs like *afwerken* (*ik werk dit af*)[12]. Other particles can be singular common nouns in standard case `N(soort,ev,stan)`, such as in verbs like *paardrijden* (*ik rij graag paard*)[13]. Yet another set of particles can be tagged as adverbially used adjectives `ADJ(vrij)`, such as in *vrijspreken* (*de rechter sprak hem vrij*).[14] A final category of particles are the real adverbs `BW`, such as in *bijeenbrengen* (*hij brengt geld bijeen*).[15] Each of the words of a sentence that is tagged as a verb, be it in its finite, infinite or past participle form is combined with each of the words tagged with one of the potential particles. The most likely combination according to an 80 million word corpus is selected, merging the verb with its particle. This procedure is recursively applied until no further separable verbs are detected. We apply the compounding approach of Vandeghinste (2002) to separable verb detection.

Then we apply lemmatization. Apart from the lexicon lookup procedure, we have implemented a rule-based lemmatizer that uses the token and part-of-speech tag as input and returns the lemma. The rules consist of a number of regular expression substitutions, depending on the part-of-speech tag of the input token.

4.2. Semantic analysis

As a first step in the semantic analysis of the source message, we detect words indicating a negative polarity, such as *niet* (not) and *geen* (no). When such a word is found, we look for its head. In the case of *niet*, we look for a verb within a window size of 3. When such a verb is found, then we add the value negative to the word feature polarity.

As a second step in semantic analysis, we look up all the possible Cornetto-synsets connected to the lemma of each word. We filter these synsets, keeping those where the part-of-speech of the synset agrees with the part-of-speech main category of the word, as labeled by the part-of-speech tagger.

4.3. Translating into pictographs

To each of the synsets that have been attributed to the words, we attach the Sclera pictographs that are linked to these synsets. We consider different types of linking pictographs with synsets: First, we have the pictographs that are linked to one and only one synset, which we call the `sclera_single pictographs`. Secondly, we have the pictographs that represent a more complex concept than a single synset, and these pictographs have been linked to two or more synsets. For each of these complex synsets we consider one of the synsets as the head synset, and the rest as dependents.

So, for every synset of every word we distinguish three types of Sclera pictographs connected to that synset: the `sclera_single` pictographs, the `sclera_complex` pictographs for which the head synset equals the synset of the word and the `sclera_as_dependent` pictographs for which one of the dependent's synsets matches the synset of the word.

Because we expect the coverage of the Sclera pictographs to be too low for practical usage, we decided to extend the coverage of the system, by using the Cornetto relations between synsets. An overview of the Cornetto relations between synsets that are used in our system is presented in Figure 14.

`XPOS_NEAR_SYNONYM` indicates a link between similar concepts but with a different part-of-speech.

`HAS_HYPERONYM` indicates the link from a subcategory to a supercategory. We keep track of how many of these synset links we apply and count penalties for using these relations, which results in prefering the words closest to the

[12] finish (I finish this)
[13] ride a horse (I like horse riding)
[14] acquit (The judge acquited him)
[15] collect (He collects money)

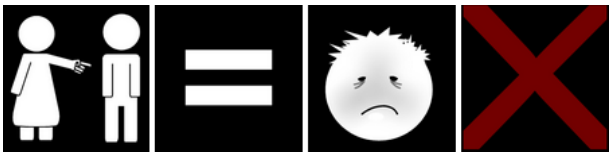

Figure 13: Translation of the message *"He is recovered"*

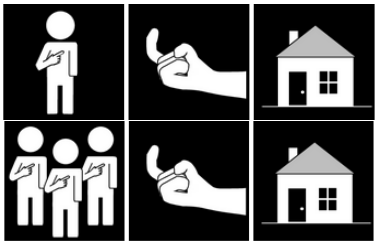

Figure 15: When handled as language (2014)

original meaning. We set `HYPERONYM PENALTY=5` and `XPOS PENALTY=2`, and we recursively apply this coverage expansion until we reach the penalty threshold, which we currently have set to 8.

In a few cases we use the `ANTONYM` relation, for example, *genezen* (recovered) is translated as *niet ziek* (not ill): there is no pictograph for *genezen*, probably because it is hard to depict this state of affairs. The translation of this message in Sclera is shown in Figure 13.[16]

Sometimes a hyponym is taken as preferred link (manually). *Confituur* (jam) is linked with the pictograph for *aardbeiconfituur* (strawberry jam). As there is no pictograph for the general concept *confituur*, the system selects a *hyperonym* as preferred translation into a pictograph. As this pictograph depicts all sweet sandwich fillings (chocolate confetti, jam, peanut butter, ...), we overruled it via the dictionary described in section 3.3., choosing a prototypical jam (in Belgium the strawberry variant) as correct translation.

The final step consists of finding the optimal path. For this we implemented an A* search algorithm (Hart et al., 1968) that finds the best path. The best path is considered as the path with the lowest combined penalty. Each pictograph that is used has a cost of one, hence preferring complex pictographs over representing multiple synsets over multiple simplex pictographs each representing a single synset.

As a result, the sentence *ik kom naar huis* is now translated as shown in Figure 15 (cf also Figure 7).

Evaluation

We have manually selected a test corpus of 50 emails that have been sent in the WAI-NOT environment. This concerns emails of which the intended message was clear to the selector, as many of the emails sent with the WAI-NOT system are clearly the result of playing around with the pictograph input interface and not of an urge to communicate.

[16]The last pictogram consists of a red cross against a black background.

	Baseline	Dutch2Sclera	Improvement
Precision	83.12%	87.33%	5.06%
Recall	28.92%	79.56%	**175.10%**
F-Score	42.91%	83.27%	94.06%

Table 3: Evaluation of the Text2Sclera convertor

Table 3 presents the results of the manual analysis of the translation quality of these email messages.

The baseline shows the accuracy of an approach in which the words are translated into pictographs if they match the filename of the pictograph (without the extension).

These results clearly show a large improvement on recall of content words (relative rise of 175%), which can be largely attributed to the presented resources. Although no word sense disambiguation algorithms are used, shallow linguistic analysis already provides a 5% relative rise in precision. There is clearly further room for improvement by applying better automatic spelling correction mechanisms as the input data is very noisy.

5. Conclusions and future work

It is clear from the evaluation that our system provides an improvement in the communication possibilities of illiterate people, although further improvements are surely possible provided more research is done. It is with this purpose that we provide the research community with the resources we created.

Future work will consist of scaling to other languages, such as English and Spanish, and other pictograph sets, such as Beta. This work started in spring 2014.

More future work needs to be done with respect to the translation engine, by including proper word sense disambiguation algorithms and better spelling correction.

Another aspect that we are working on is the translation of pictographic messages into natural language text, in order to allow bidirectional communication.

Apart from that, some more extensive evaluation is under way, including an *in vivo* evaluation of how the use of such a pictograph translation engine improves the living conditions and quality of users with IDD in a number of different settings.

6. Acknowledgements

This research is done in the Picto project, funded by the Support Fund Marguerite-Marie Delacroix.[17] Follow up work on the localisation of the text to pictograph translator is funded by the European Commission CIP-621055 in the Able-to-Include project.

7. References

Brants, T. (2000). TnT A Statistical Part-of-speech Tagger. In *Proceedings of the 6th Applied Natural Langauge Processing Conference (ANLP)*, pages 224–331, Seattle, Washington, USA. ACL.

[17]http://www.fondsmmdelacroix.org/en/

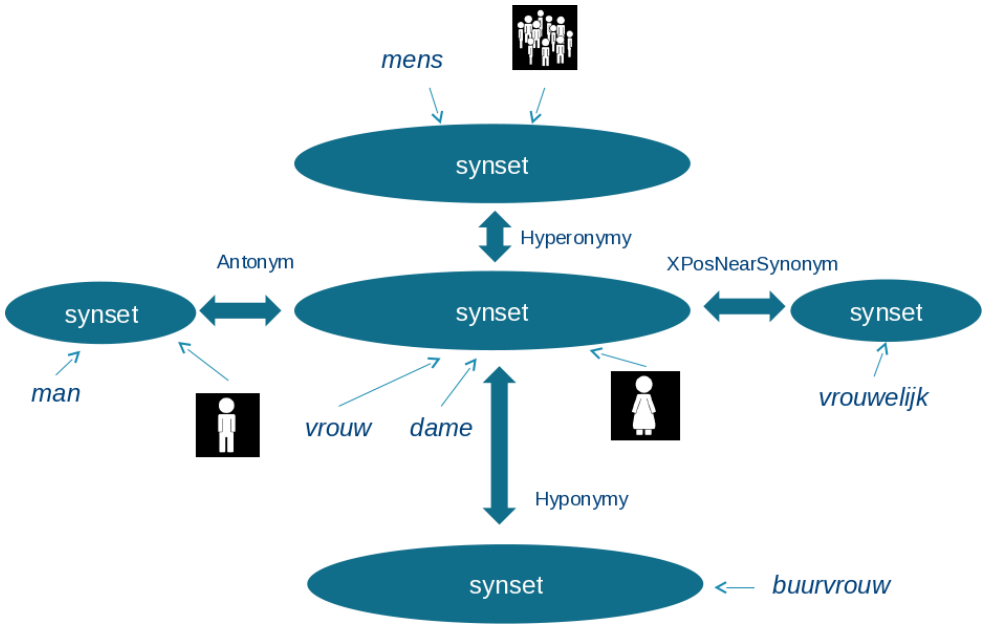

Figure 14: Cornetto, synsets and pictographs

Davies, D., Stock, S., and Wehmeyer, M. (2001). Enhancing independent internet access for individuals with mental retardation through use of a specialized web browser: a pilot study. *Education and Training in Mental Retardation and Developmental Disabilities*, 36(1):107–113.

Dawe, M. (2006). Desperately seeking simplicity: how young adults with cognitive disabilities and their families adopt assistive technologies. In *Proceedings of the SIGCHI conference on Human Factors in computing systems (CHI 06)*, pages 1143–1152, New York, NY, USA. ACM.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Goldberg, A., Zhu, X., Dyer, C., Eldawy, N., and Heng, L. (2008). Easy as ABC? Facilitating Pictorial Communication via Semantically Enhanced Layout. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 119–126. ACL.

Halcsy, P., Kornai, A., and Oravecz, C. (2007). HunPos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 209–212, Prague, Czech Republic. ACL.

Hart, P. E., Nilsson, N. J., and Raphael, B. (1968). A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics SSC4*, 4(2):100–107.

Joshi, D., Wang, J., and Li, J. (2006). The Story Picturing Engine A System for Automatic Text Illustration. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1):1–22.

Keskinen, T., Heimonen, T., Turunen, M., Rajaniemi, J., and Kauppinen, S. (2012). The Story Picturing Engine A System for Automatic Text Illustration. *Interacting with Computers*, 24(5):374–386.

Mihalcea, R. and Leong, C. (2008). Toward communicating simple sentences using pictorial representations. *Machine Translation*, 22(3):153–173.

Miller, G. (1995). Wordnet: A lexical database. *Communications of the ACM*, 38(11):39–41.

Newell, A., Carmichael, A., Gregor, P., and Alm, N. (2002). Information technology for cognitive support. In Jacko, J. A. and Sears, A., editors, *The Human-Computer Interaction Handbook*, pages 464–481, Hillsdale, N.J., USA. Lawrence Erlbaum Associates.

Oostdijk, N., Reynaert, M., Hoste, V., and Schuurman, I. (2013). The Construction of a 500-million-word Reference Corpus of Contemporary Written Dutch. In Spyns, P. and Odijk, J., editors, *Essential Speech and Language Technology for Dutch: resources, tools and applications*. Springer.

Van Eynde, F., (2005). *Part of speech tagging en lemmatisering van het D-CoI corpus*.

Vandeghinste, V. (2002). Lexicon Optimization: Maximizing Lexical Coverage in Speech Recognition through Automated Compounding. In Rodrguez, M. and Araujo, C., editors, *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, pages 224–331, Las Palmas, Spain. ELRA.

Vandeghinste, V. (2012). Bridging the Gap between Pictographs and Natural Language. In *Proceedings of the W3C/WAI Online Symposium: Easy-to-Read on the Web*. w3.org.